UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P O Box 1450
Alexandria, Virginia 22313-1450
www uspto gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/788,455	03/01/2004	Kent Bodell	C697 0007/GNM	7385

720 7590 12/17/2007
OYEN, WIGGS, GREEN & MUTALA LLP
480 - THE STATION
601 WEST CORDOVA STREET
VANCOUVER, BC V6B 1G1
CANADA

EXAMINER	
GUPTA, MUKTESH G	

ART UNIT	PAPER NUMBER
4121	

MAIL DATE	DELIVERY MODE
12/17/2007	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Office Action Summary	**Application No.** 10/788,455	**Applicant(s)** BODELL ET AL.	
	Examiner Muktesh G. Gupta	**Art Unit** 4121	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE *3* MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on *01 March 2004*.

2a)□ This action is **FINAL**. 2b)☒ This action is non-final.

3)□ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1-20* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)□ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *1-20* is/are rejected.

7)□ Claim(s) _____ is/are objected to.

8)□ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)□ The specification is objected to by the Examiner.

10)☒ The drawing(s) filed on _____ is/are: a)☒ accepted or b)□ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)□ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)□ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)□ All b)□ Some * c)□ None of:

1.□ Certified copies of the priority documents have been received.

2.□ Certified copies of the priority documents have been received in Application No. _____.

3.□ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)

2) □ Notice of Draftsperson's Patent Drawing Review (PTO-948)

3) ☒ Information Disclosure Statement(s) (PTO/SB/08)
Paper No(s)/Mail Date *04/04/2005*.

4) □ Interview Summary (PTO-413)
Paper No(s)/Mail Date. _____ .

5) □ Notice of Informal Patent Application

6) □ Other: _____.

DETAILED ACTION

1. **Claims 1-20** have been examined and are pending.

Claim Rejections - 35 USC § 102

The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

> A person shall be entitled to a patent unless –
>
> (a) the invention was known or used by others in this country, or patented or described in a printed publication in this or a foreign country, before the invention thereof by the applicant for a patent.

2. **Claims 1-20** are rejected under 35 U.S.C. 102(a) as being anticipated by Non-Patent Publication titled "Performance Modeling of Message-passing Parallel Program" to Grove, Duncan A., (hereinafter "Grove").

As to Claim 1, Grove anticipates method for communicating data from a first compute node of a computer system comprising multiple compute nodes interconnected by an inter-node communication network to a second one of the multiple compute nodes, the method comprising (as stated in Page 44, section 2.21, lines 19-21, ***network*** of symmetrical multiprocessors, SMP ***nodes***, each with their own local memory and one or more processors. Processors are ***connected*** by multiple levels of ***bus-based communication links***):

placing the data on a full-duplex packetized interconnect directly connecting a CPU of the first compute node to a network interface connected to the inter-node communication network (as stated in Page 143, section 4.6, lines 32-34, page 144, section 4.6, lines 1-2, MPI ***send receive operation*** involves the simultaneous exchange of two messages between a ***pair*** of ***processes***, ***one in each direction***. Physically, many machines including all three of the machines benchmarked here have ***full-duplex network links*** so there is the potential for MPI ***send receive*** messages to actually ***transit*** those ***interconnect networks*** simultaneously);

receiving the data at the network interface (as stated in preceding lines and page 66, section 3, lines 30-32, ***network interface*** card will be instructed to fetch the ***data*** from wherever it is in memory and to place it on to the ***external communication network***);

and, transmitting the data to a network interface of the second compute node by way of the inter-node communication network (as stated in preceding lines, and page 66, section 3, lines 5-7, ***send receive*** messages to actually ***transit*** those ***interconnect networks*** simultaneously. Once the message arrives at the destination host, it follows the reverse chain of events that occurred at the source, and it finally arrives for the user program to process).

As to Claim 2, Grove anticipates method according to claim 1 wherein the network interface and the CPU are the only devices configured to place data on the packetized interconnect (as stated in page 63, section 3, lines 29-32 and page 118, section 4, lines 27-29, lines 35-37, it is going to be assumed from here on that no local computation involves significant disk I/O. This also precludes access to virtual memory, i.e. more than enough core physical memory is assumed to be available. ***Communications processor*** on each QsNet ***network interface*** can do a substantial amount of the work required by higher level protocols such as MPI without the ***intervention*** of a ***node's*** main ***CPU***. Portable Batch System (PBS) that the APAC NF uses made it possible to gain ***dedicated*** access to a partition of consecutive ***nodes*** on the machine and ***interconnect***).

As to Claim 3, Grove anticipates method according to claim 1 comprising transmitting the data from the network interface to the second computer node by way of a full-duplex communication link of the inter-node communication network (as stated in page 63, section 3, lines 15-17, lines 29-32 and page 118, section 4, lines 27-29, lines 35-37, page 144, section 4, lines 11-13, Portable Batch System (PBS) that the APAC NF used made it possible to gain ***dedicated*** access to a partition of consecutive ***nodes*** on the machine and ***interconnect.*** The code had dedicated access to the ***CPU*** and memory and characterized situation where the ***full-duplex*** nature of the ***inter-connect network*** is able to support the simultaneous ***communication*** of the MPI Sendrecv operation without much performance degradation).

As to Claim 4, Grove anticipates method according to claim 3 comprising passing the data through a buffer at the network interface before transmitting the data (as stated in preceding paragraphs and page 66, section 3, lines 26-32, At any of points during transmission through network system, there is the potential for the ***message (data)*** to be ***buffered***, and possibly for control to be returned to the calling process, while the ***message (data)*** is dealt with asynchronously. Once the ***message (data)*** has been delivered to the operating system, it will undergo further processing by the network protocol stack. At this point the ***network interface*** card will be instructed to fetch the ***data*** from wherever it is in ***memory*** and to place it on to the ***external communication network***).

As to Claims 5 and 19, Grove anticipates method and compute node according to claims 1 and 11, comprising, at the network interface, determining a size of the data and, based upon the size of the data, selecting among two or more protocols for transmitting the data (as stated in preceding paragraphs and page 131, section 4, lines 7-12, For ***messages*** up to ***16 Kbytes*** minus 96 or 112 book-keeping bytes for 32-bit or 64-bit applications respectively, GM uses an ***eager*** message delivery ***protocol*** where messages are always sent immediately and ***buffered*** by the MPI implementation at the receiver. ***Messages larger*** than this but less than 32 Kbytes minus 8 bytes are sent using a ***rendezvous protocol***, where the sender does not transmit any data until the receiver acknowledges that it has buffer space to receive it).

As to Claim 6, Grove anticipates method according to claim 5 wherein the two or more protocols comprise an eager protocol and a rendezvous protocol (as stated in preceding paragraphs and page 131, section 4, lines 7-12, For ***messages*** up to ***16 Kbytes*** minus 96 or 112 book-keeping bytes for 32-bit or 64-bit applications respectively, GM uses an ***eager message*** delivery ***protocol*** where messages are always sent immediately and ***buffered*** by the MPI implementation at the receiver. ***Messages larger*** than this but less than 32 Kbytes minus 8 bytes are sent using a ***rendezvous protocol***, where the sender does not transmit any data until the receiver acknowledges that it has buffer space to receive it).

As to Claim 7, Grove anticipates method according to claim 6 comprising, upon selecting the rendezvous protocol, automatically generating a Ready To Send message at the network interface of the first compute node (as stated in preceding paragraphs and page 131, section 4, lines 7-12, For ***messages*** up to ***16 Kbytes*** minus 96 or 112 book-keeping bytes for 32-bit or 64-bit applications respectively, GM uses an ***eager*** message delivery ***protocol*** where messages are always sent immediately and ***buffered*** by the MPI implementation at the receiver. ***Messages larger*** than this but less than 32 Kbytes minus 8 bytes are sent using a ***rendezvous protocol***, where the ***sender*** does not ***transmit*** any ***data*** until the ***receiver acknowledges*** that it has ***buffer space*** to ***receive*** it).

As to Claims 8 and 15, Grove anticipates method and compute node according to claims 1 and 11, wherein the data comprises a raw ethertype datagram and transmitting the data comprises encapsulating the raw ethertype datagram within one or more link layer packet headers (as stated in preceding paragraphs and page 162, section 4, lines 12-15 and lines 21-26, page 163, lines 15-20, in ***TCP/IP/Fast Ethernet*** combination, substantial numbers of outliers in message-passing times were observed on Perseus. For transmission over a physical network, each internet-layer IP ***datagram*** must be split into a number of ***link-layer packets***, each up to Maximum Transmission Unit (MTU) bytes in size, and ***encapsulated*** along with a ***link-layer header*** into what is called a ***frame***. For example, in the case of ***MPI/TCP/IP/Ethernet*** communication, MPI messages are eventually split into a number of Ethernet frames, which are transmitted through the network and reassembled at the destination host. In order to provide reliable communication, ***TCP/IP*** requires that all ***data messages*** also known as segments in TCP/IP parlance must be acknowledged by the receiver, so that the sender is aware of their safe delivery. To facilitate this, every ***data message*** is ***augmented*** with an ***ordinal sequence number*** and some other information, the ***complete package*** is called an ***IP datagram***. Then, upon receipt of a data message, the receiver sends a special acknowledgement message containing that sequence number back to the original sender).

As to Claim 9, Grove anticipates method according to claim 8 wherein the link layer packet headers comprise InfiniBand.TM. link layer packet headers (as stated in preceding paragraphs and page 117, section 4, lines 28-34, page 118, lines 1-8, Orion nodes are connected by both 100 Mbit/s Ethernet and a Myrinet network, which provides 1.28 Gbit/s of full duplex bandwidth and a significantly lower latency than Fast Ethernet. The Myrinet hardware provides an inherently ***connection-less data transfer mechanism***, on top of which is built a ***GM protocol layer*** which provides ***reliable and ordered end-to-end packet delivery to user-space processes.*** Orion has a peak speed of 144 Gflop/s with a Linpack benchmark result of 110 Gflop/s).

As to Claim 10, Grove anticipates method according to claim 1 wherein the data comprises a raw internet protocol datagram and transmitting the data comprises encapsulating the internet protocol datagram within one or more link layer packet headers (as stated in preceding paragraphs and page 162, section 4, lines 12-15 and lines 21-26, page 163, lines 15-20, in ***TCP/IP/Fast Ethernet*** combination, substantial numbers of outliers in message-passing times were observed on Perseus. For transmission over a physical network, each ***internet-layer IP datagram*** must be split into a number of ***link-layer packets***, each up to Maximum Transmission Unit (MTU) bytes in size, and ***encapsulated*** along with a ***link-layer header*** into what is called a ***frame).***

As to Claim 11, Grove anticipates compute node for use in a multi-compute-node computer system (as stated in Page 3, section 1.2, lines 31-34, A popular contemporary trend in constructing parallel computers is to connect ***large numbers***

of ***SMP nodes*** using ***multicomputer communication networks***, in an attempt to get the best of both systems);

the compute node comprising:

a CPU (as stated in Page 63, section 3, lines 15-17, Assuming that the code has ***dedicated*** access to the ***CPU*** and memory which is a reasonably accurate assumption provided that no other user programs are running on the CPU);

a network interface (as stated in preceding lines and page 66, section 3, lines 30-32, ***network interface*** card will be instructed to fetch the ***data*** from wherever it is in ***memory*** and to place it on to the ***external communication network***);

and, a dedicated full-duplex packetized interconnect directly coupling the CPU to the network interface (as stated in Page 143, section 4.6, lines 32-34, page 144, section 4.6, lines 1-2, Physically, many machines including all three of the machines benchmarked here have ***full-duplex network links*** so there is the potential for MPI ***send receive*** messages to actually ***transit*** those ***interconnect networks*** simultaneously).

As to Claim 12, Grove anticipates compute node according to claim 11 wherein the dedicated packetized full-duplex interconnect is not shared by any devices other than the CPU and the network interface (as stated in page 63, section 3, lines 15-17, lines 29-32 and page 118, section 4, lines 27-29, lines 35-37, page 144, section 4, lines 11-13, Portable Batch System (PBS) that the APAC NF used made it possible to gain ***dedicated*** access to a partition of consecutive ***nodes*** on the machine and

interconnect. The code had dedicated access to the ***CPU*** and ***memory*** and characterized situation where the ***full-duplex*** nature of the ***inter-connect network*** is able to support the simultaneous ***communication*** of the MPI Sendrecv operation without much performance degradation).

As to Claim 13, Grove anticipates compute node according to claim 11 comprising a memory, and a facility configured to allocate eager protocol buffers in the memory and to automatically signal to one or more other compute nodes that the eager protocol buffers have been allocated (as stated in preceding lines and page 66, section 3, lines 30-32, and page 131, section 4, lines 7-12, For ***messages*** up to ***16 Kbytes*** minus 96 or 112 book-keeping bytes for 32-bit or 64-bit applications respectively, GM uses an ***eager*** message delivery ***protocol*** where messages are always sent immediately and ***buffered*** by the ***MPI*** implementation at the ***receiver***. Network ***interface*** card will be instructed to fetch the ***data*** from wherever it is in ***memory*** and to place it on to the ***external communication network***).

As to Claim 14, Grove anticipates compute node according to claim 13 comprising a facility configured to automatically associate memory protection keys with the eager protocol buffers and a facility configured to verify memory protection keys in incoming eager protocol messages before writing the incoming eager protocol messages to the eager protocol buffers. (as stated in preceding paragraphs and page 131, section 4, lines 7-12 and page 118, section 4, lines 23-26,

messages up to ***16 Kbytes*** minus 96 or 112 book-keeping bytes for 32-bit or 64-bit applications respectively, GM uses an ***eager message*** delivery ***protocol*** where messages are always sent immediately and ***buffered*** by the MPI implementation at the receiver, where the sender of a packet always waits for a subsequent end of packet ***token*** from the receiver before it completes. The link protocol detects any lost messages or network faults and organizes ***retransmission*** and re-routing as necessary).

As to Claim 16, Grove anticipates compute node according to claim 11 wherein the network interface comprises a buffer connected to buffer outgoing data (as stated in preceding lines and page 66, section 3, lines 30-32 and page 207, section 5, lines 13-18 ***network interface*** card will be instructed to fetch the ***data*** from wherever it is in ***memory*** and packets and more data accumulated in the switches' buffers, the switches had to undertake more processing to determine which packets should be scheduled for output first and to place it on to the ***external communication network***).

As to Claim 17, Grove anticipates compute node according to claim 11 comprising a plurality of CPUs each connected to the interface by a separate dedicated full-duplex packetized interconnect. (as stated in preceding lines and page 44, section 2.21, lines 19-21, page 143, section 4.6, lines 32-34, page 144, section 4.6, lines 1-2 network of SMP ***nodes***, each with their ***own*** local ***memory*** and ***one*** or

more processors. Processors are connected by multiple levels of bus-based communication links have ***full-duplex network links*** so there is the potential for MPI Sendrecv messages to actually transit those ***interconnect networks*** simultaneously).

As to Claim 18, Grove anticipates compute node according to claim 11 wherein the CPU is connected to each of a plurality of network interfaces by a plurality of dedicated full-duplex packetized interconnects (as stated in preceding lines and page 5, section 1.2, lines 10-12, page 143, section 4.6, lines 32-34, page 144, section 4.6, lines 1-2, alternative to buffering messages in the network is to use processors in ***nodes*** with ***multiple network interfaces*** as routing intermediaries. Physically, many machines have ***full-duplex network*** links so there is the potential for MPI Sendrecv messages to actually transit those ***interconnect networks*** simultaneously).

As to Claim 20, Grove anticipates computer system comprising a plurality of compute nodes according to claim 11 interconnected by an inter-node data communication network, the inter-node data communication network providing at least one full-duplex data link to the network interface of each of the nodes (as stated in preceding lines and page 117, section 4, lines 27-33, page 118, section 4, lines 1-8, Orion is a Sun Technical Compute Farm at the University of Adelaide. It consists of a ***cluster*** of 40 Sun E420R SMP servers, each with four 450MHz

UltraSparc II ***processors*** and 4GB of RAM. The ***nodes*** are connected by both ***100 Mbit/s Ethernet*** and a Myrinet network, which provides 1.28 Gbit/s of ***full duplex*** bandwidth and a significantly lower latency than Fast Ethernet. The Myrinet hardware provides an inherently ***connection-less data transfer mechanism***, on top of which is built a GM protocol layer which provides reliable and ordered end-to-end packet delivery to user-space processes with a Linpack benchmark result of 110 Gflop/s).

Conclusion

3. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure.

US Patent Publication No. 20040103218 to Blumrich, Matthias A et al., US Patent Publication No. US 20010034798 to Reed, Coke S, US Patent No. 7093024 to Craddock, David F. et al., and US Patent No. 7155537 to Weber, Bret S. et al., are cited for reference.

Any inquiry concerning this communication or earlier communications from the examiner should be directed to MUKTESH G. GUPTA whose telephone number is (571)270-5011. The examiner can normally be reached on Monday-Friday, 8:00 a.m. -5:00 p.m., EST.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Taghi T. Arani can be reached on 571-272-3787. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

MG

/Taghi T. Arani/
Supervisory Patent Examiner, Art Unit 4121
12/13/2007